The EMBO Journal vol.3 no.11 pp.2491 – 2497, 1984

The DNA sequence analysis of soybean heat-shock genes and identification of possible regulatory promoter elements

Fritz Schöffl, Eberhard Raschke and Ronald T.Nagao[1]

Universität Bielefeld, Fakultät für Biologie (Genetik), D-4800 Bielefeld 1, FRG and [1]University of Georgia, Botany Department, Athens, GA 30602, USA

Communicated by B.Jockusch

The soybean possesses a gene family encoding the major low mol. wt. heat-shock proteins of 15 – 18 kd. We have determined the primary DNA sequences of two of the genes, both located on the same subgenomic DNA fragment. The protein coding regions are characterized by long uninterrupted open reading frames and by sequence homology of 92% and 100% with a heat-shock specific cDNA. One protein sequence deduced from the completely cloned gene *hs6871* is composed of 153 amino acids with a total mol. wt. of 17.3 kd; the other protein is a truncated polypeptide containing 73 amino acids at the carboxy-terminal end of an incompletely cloned heat-shock gene designated *hs6834*. Investigations of the hydrophilic/hydrophobic characteristics of the polypeptides revealed a conservation of structural features between heat-shock proteins from soybean, *Caenorhabditis* and *Drosophila* and mammalian lens α-crystallin. The 5′ end of the soybean heat-shock gene *hs6871* was mapped by S1 nuclease at a position which is ~100 nucleotides upstream from the translation start codon and 25 nucleotides downstream from a TATA-box sequence. Six other potential promoter elements which are homologous to the *Drosophila* heat-shock consensus sequence CT-GAA–TTC-AG-, are present within ~150 nucleotides upstream from the TATA-box. The possible functions of these promoter elements in transcriptional regulation of expression of soybean heat-shock gene are discussed.

Key words: DNA sequences/*hs* genes/promoter elements/soybean/S1 mapping

Introduction

High temperature stress or heat-shock (hs) induces the vigorous synthesis of several new proteins, the heat-shock proteins (hsps) in a variety of species. The synthesis of most other proteins is concomitantly reduced by transcriptional and translational control mechanisms (for review, see Ashburner and Bonner, 1979; Schlesinger *et al.*, 1982). The hs response is thought to contribute to homeostasis with the hsps having a protective role presumed to counteract or prevent deleterious effects induced by the heat-shock (Ashburner and Bonner, 1979; Velasquez and Lindquist, 1984; Schöffl *et al.*, 1984). The mechanism of this response must have been highly conserved during evolution, since cloned *Drosophila hs* genes for hsp 70 are thermoinducibly expressed when introduced in mouse (Corces *et al.*, 1981), monkey (Mirault *et al.*, 1982; Pelham, 1982), rat (Burke and Ish-Horowicz, 1982) and *Xenopus* cells (Voellmy and Runngger, 1982; Bienz and Pelham, 1982).

A comparison between the much studied *Drosophila* system and the soybean hs response as a model system for higher plants (Schöffl *et al.*, 1984; Key *et al.*, 1984), reveals hsps of similar sizes induced in *Drosophila* at 37°C, and in soybean at 40°C (Key *et al.*, 1981). However, there are also marked differences between these two organisms in the mol. wt., complexity and abundance of a group of low mol. wt. hsps. In *Drosophila*, a chromosomal cluster of four genes encodes the hsps 22, 23, 26 and 27 kd (Corces *et al.*, 1980; Wadsworth *et al.*, 1980; Voellmy *et al.*, 1981) but more than 20 hsps with mol. wts. of 15 – 18 kd are synthesized in soybean (Key *et al.*, 1981; Schöffl and Key, 1982). These hsps are encoded by multigenic families in soybean (Schöffl and Key, 1983), and they are abundantly expressed during hs, averaging 20 000 poly(A) mRNA molecules per gene in the cell (Schöffl and Key, 1982). cDNA clones isolated from hs transcripts (Schöffl and Key, 1982) cross-hybridize with hs-specific mRNAs of a variety of plant species (Key *et al.*, 1983; Schöffl, 1984). All the plants examined to date also synthesize low mol. wt. hsps, suggesting a common role for these proteins in plant hs response (Key *et al.*, 1983; Schöffl *et al.*, 1984). We have studied the regulation of transcription of plant hs genes, and in particular members of the plant-specific class I multigene family, since little is known about plant gene structure and regulation in general and nothing about plant hs genes.

Many of the *Drosophila hs* genes have already been cloned and sequenced. Various investigators have shown that striking homologies exist between the four hs-activated genes encoding hsp 22, 23, 26 and 27, which are located on a 11-kb DNA fragment in chromosomal subdivision 67B (Ingolia and Craig, 1982; Southgate *et al.*, 1983). Each of these genes also possesses characteristic eucaryotic 5′ and 3′ sequence elements. Several copies of the most highly conserved hsp of *Drosophila* (hsp 70) have been sequenced (Karch *et al.*, 1981; Ingolia *et al.*, 1980); they are closely homologous to each other and the homology extends for several hundred base pairs on the 5′ side of the transcribed gene region. Expression of cloned hsp 70 genes in heterologous systems controlled by hs is regulated by an essential region upstream from the TATA-box that bears sequence homology to other hs promoters (Pelham, 1982). The so-called 'hs consensus sequence' 5′*CT*g*GAA*tn*TTC*t*AG*a is active in controlling hs-regulated expression of foreign genes even when synthetic promoter elements are used which have only 80% homology with the symmetric consensus sequence (Pelham and Bienz, 1982).

Previously, we have cloned and characterized a genomic DNA fragment containing two soybean hs genes of the class I multigene family (Schöffl and Key, 1983). One of these genes, designated *hs6871*, was precisely mapped by electron microscopic R-loop analysis, revealing a short coding region of 400 – 500 bp without detectable intervening sequences. This paper reports the DNA sequence of the entire gene coding region and of its 5′- and 3′-flanking sequences. We have also sequenced an incomplete second gene, present on the same genomic clone, and its homologous cDNA. The data have been used to predict the amino acid sequences and to examine

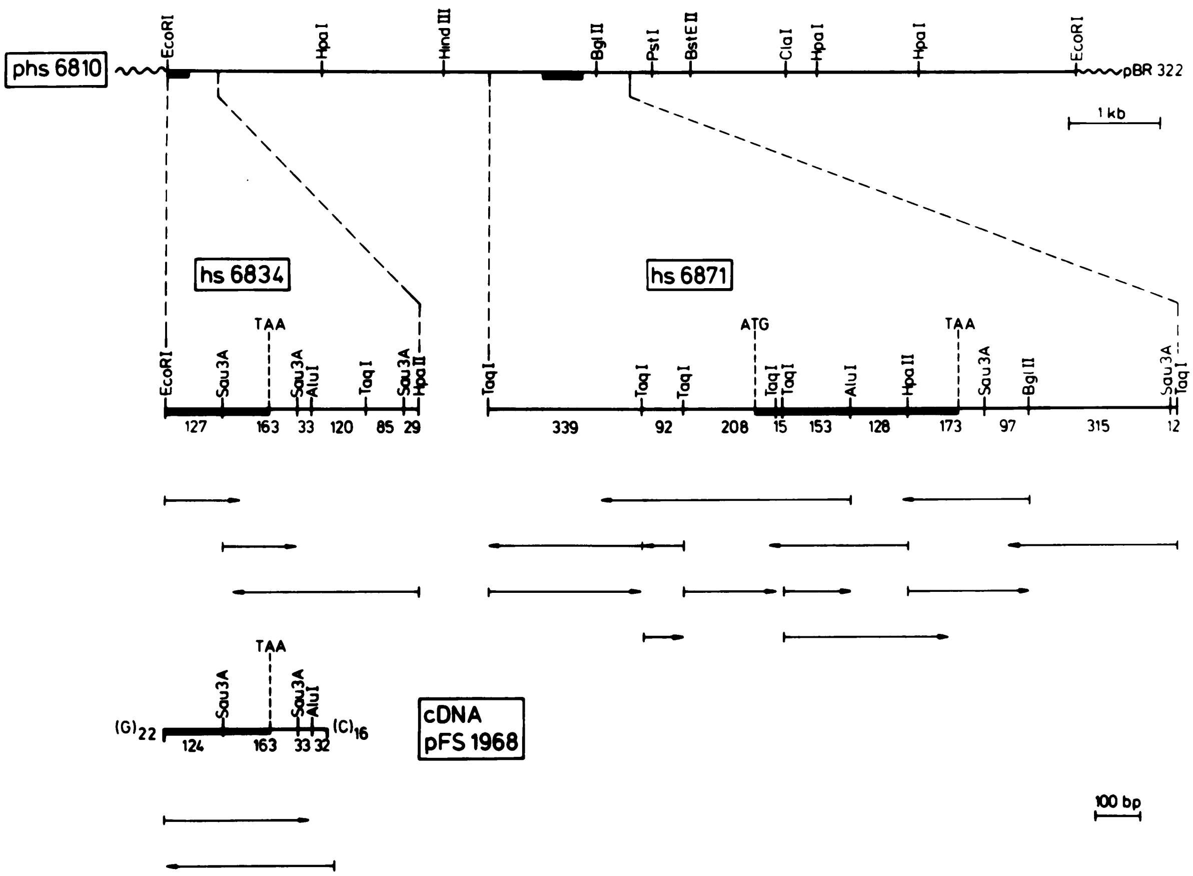

Fig. 1. Map of restriction sites of a genomic DNA segment and of a cDNA clone. The sequencing strategy is outlined by arrows. Filled parts indicate protein coding regions, fragment sizes are in nucleotides.

structural features in comparison with hs proteins from other organisms and to identify possible regulatory promoter elements controlling transcription.

Results

The overall organization of hs genes and restriction endonuclease sites on a 10.5-kb genomic DNA fragment (subclone *phs6810*; Schöffl and Key, 1983) is shown in the upper part of Figure 1. The approximate location of the two different *hs* genes, *hs6834* and *hs6871* was determined by 'Southern blot' hybridization (Southern, 1975) with the hs-specific class I cDNA insert of pFS1968 (data not shown). The exact position and the orientation of gene *hs6871* were mapped previously by electron microscopic R-loop analysis (Schöffl and Key, 1983). The strategy used to determine the DNA sequence of the two *hs* genes and of the corresponding cDNA is outlined in the lower part of Figure 1.

Nucleotide sequence of hs6871

The nucleotide sequence of the entire coding region of *hs6871*, its promoter, the 5′ upstream region and 50% of the 3′-flanking sequences was determined for both strands. A continuous DNA stretch of 1537 nucleotides was sequenced and analysed (Figure 2). There is only one large open reading frame present, starting at nucleotide position + 1 with ATG and translating into a protein of 153 amino acid residues, or 17.3 kd, terminated by a TAA nonsense codon after 459 bp. The ATG start codon is preceded in the same reading frame by two nonsense codons at position − 12 (TGA) and − 9 (TAA). Positive nucleotide numbers start with + 1 at the beginning of the gene coding region and proceed in the 3′ direction. Negative numbers refer to the 5′ upstream region of *hs6871*, starting with − 1 lefthand to the ATG initiation codon of the protein sequence in Figure 2. All other potential protein coding regions are interrupted after < 150 − 200 nucleotides in all three reading frames and in both orientations. The location, the polarity and the size of the identified hsp coding region are in good agreement with the R-loop and restriction mapping data and with the size of class I hsps (Schöffl and Key, 1982, 1983). This coding region is clearly distinguished by its GC content of 47% compared with only ~30% in its 3′- and 5′-flanking regions. The amino acid composition of the protein and the codon utilization of *hs6871* are shown in Table I.

Nucleotide sequences of hs6834 and its corresponding cDNA clone pFS1968

The 490-bp *Pst*I insert of pFS1968 was sequenced in both strands determining the entire sequence of the cDNA (see Figure 1). According to the cDNA cloning strategy used for hs poly(A) mRNA (Schöffl and Key, 1982), GC tails (22 bp and 16 bp) are present at the ends of the cDNA (see Figure 1). The sequence of the internal 352 nucleotides of the cDNA is identical with the DNA sequence of the genomic hs gene *hs6834* (Figure 2). There is a long open reading frame, coming in from the 5′ end with unknown origin, translating into a truncated polypeptide of 73 amino acid residues and terminated by the nonsense codon TAA. This coding region is highly homologous with the 3′ half of *hs6871*. The numbering of nucleotides and amino acids refers to the homologous

hs6871

```
5'    TCGAGAAAAAAAATTCATTATATTATTGATATAAAATATTCATTAATTTTATCAATAATTAATTTATATTTATATTGAGAAATCTAGATAGTCA

-500  GCCTTTTAAGAGATAGAATTTAAAATATAATTTGCGTAAAACATTATTAAAAATACAAATTTATAAATTAAGTTCAACTCATCCTATCTCACTCTTTAAA

-400  TACGATGTTTACTTATTAGACTCATTAATAAAAAAAAAAAAAATCATTTGTACAAAGCCCACCATAAAGGCAATTTGGGCCTGGTAGACCAATCCTAACC

-300  AATGTCTGGTTAAGATGGTCCAATCCCGAAACTTCTAGTTGCGGTTCGAAGAAGTCCAGAATGTTTCTGAAAGTTTCAGAAAATTCTAGTTTTGAGATTT

-200  TCAGAAGTACGGCATGATGATGCATAACAAGGACTTTCTCGAAAGTACTATATTGCTCCTCTACATCATTTTAAATACCCCATGTGTCCTTTGAAGACAC

-100  ATCACAGAAAGAAGTGAAGGCATCGTTAGCAGTTTTGTAGATTCAACCTCAATTTGCAGAGTTACGTTCTAATATATTTACACAAGACTGATAAGAGAAA

                                       10                                      20
Met Ser Leu Ile Pro Ser Phe Phe Gly Gly Arg Arg Ser Ser Val Phe Asp Pro Phe Ser Leu Asp Val Trp Asp
ATG TCT CTG ATT CCA AGT TTC TTC GGT GGC CGA AGG AGC AGT GTT TTC GAC CCT TTC TCC CTC GAT GTG TGG GAC   + 75

Pro Phe Lys Asp Phe Pro Phe Pro Ser Ser Leu Ser Ala Glu Asn Ser Ala Phe Val Ser Thr Arg Val Asp Trp
CCC TTC AAG GAT TTT CCA TTT CCC AGT TCT CTT TCT GCT GAA AAT TCA GCG TTT GTG AGC ACA CGA GTG GAT TGG   +150

                                       60                                      70
Lys Glu Thr Pro Glu Ala His Val Phe Lys Ala Asp Ile Pro Gly Leu Lys Lys Glu Glu Val Lys Leu Glu Ile
AAG GAG ACA CCA GAA GCA CAC GTG TTC AAG GCT GAT ATT CCA GGG CTG AAG AAG GAG GAA GTG AAG CTG GAG ATT   +225

                    80                                      90                                     100
Gln Asp Gly Arg Val Leu Gln Ile Ser Gly Glu Arg Asn Val Glu Lys Glu Asp Lys Asn Asp Thr Trp His Arg
CAA GAT GGC AGA GTT CTT CAG ATA AGC GGA GAG AGG AAT GTT GAA AAA GAA GAC AAG AAT GAT ACG TGG CAT CGC   +300

                                      110                                     120
Val Glu Arg Ser Ser Gly Lys Leu Val Arg Arg Phe Arg Leu Pro Glu Asn Ala Lys Val Asp Gln Val Lys Ala
GTG GAG CGA AGC AGT GGC AAG TTG GTG AGG AGG TTT AGA TTG CCG GAG AAT GCT AAA GTG GAC CAA GTG AAG GCT   +375

                    130                                     140                                     150
Ser Met Glu Asn Gly Val Leu Thr Val Thr Val Pro Lys Glu Glu Ile Lys Lys Pro Asp Val Lys Ala Ile Asp
TCC ATG GAA AAT GGG GTT CTC ACT GTA ACT GTT CCT AAG GAA GAG ATT AAG AAG CCT GAT GTT AAG GCC ATA GAC   +450

        153
Ile Ser Gly OCHRE
ATC TCT GGT TAA      TCTATGTTGCTCTGTTCCTTCGTTGAAATGTGTTTATGTTTTCTTATTCTGAGGATCATTTGTGTGAGTCGTGTGAAA   +540

AATATTTCAGGTTTTATGTTGGCTAAGAGGCCTAATGTTTGGGCCCTAGAAATCTCTGGTTAAACTGTGTAAAGATCTGTTACTTGGTTTAAAGTTTGTG   +640

TGTTTTGTTCACTTCCAAGGAATTTATGTGTGCAAGAAAGATGTAATTGAAAAATTTAGCAATAGACTAATGGTTTTATATATTCTATGTTGCAATAAAT   +740

CTTAGGATATGTATATCACTGGAACAGATTCACTATGCCAGTGTGTGAGAAAGCAATGATAGTTCTAAATCCTCCCAGTCTACTATGCCAATGTTTTTAT   +840

ATTTTTAATTAATATTTTTTATGATGCAATAAGAAAATTAATGAGACTTTAATAAGAATAAGAATATATAACAGTCTCAACTAGCATGATCCAACAGCATCGA 3'
```

hs6834/pFS 1968

```
                    80                                      90                                     100
                    Ile Leu Gln Ile Ser Gly Glu Arg Asn Val Glu Lys Glu Asp Lys Asn Asp Thr Trp His Arg
5'               GA ATT CTT CAG ATA AGT GGA GAG AGG AAC GTT GAG AAG GAA GAC AAG AAC GAC ACG TGG CAC CGC   +300
                    |← hs cDNA clone pFS1968 →
                                      110                                     120
Val Glu Arg Ser Ser Gly Lys Phe Met Arg Ser Phe Arg Leu Pro Asp Asn Ala Lys Val Asp Gln Val Lys Ala
GTG GAG CGA AGC AGT GGT AAG TTC ATG AGG AGT TTC AGA TTG CCA GAT AAT GCT AAA GTG GAT CAA GTT AAG GCT   +375

                    130                                     140                                     150
Ser Met Glu Asn Gly Val Leu Thr Val Thr Val Pro Lys Glu Glu Ile Lys Lys Pro Asp Val Lys Ala Ile Glu
TCC ATG GAA AAT GGG GTT CTC ACT GTA ACT GTT CCA AAG GAA GAG ATT AAG AAG CCT GAT GTT AAG GCC ATA GAA   +450

        153
Ile Ser Gly OCHRE
ATT TCT GGT TAA      ACTATGTTGCTCAGTTTCTTCGTTATTGAAAAGTCGTGTGTTTATGTTTTCTTATTCTGAGGATCATTTGTATGAGTC   +540

GTGTAAAAAATATGTCAGCTATTATGTTGGTTAAGACTTAAGAAGCCTGATTATGTTAGGGCAATACAATGATACAAATCTCTGGTTAAACTGTGTTATC   +640
                     ← hs cDNA clone pFS1968 →|
TGTTACTTGGTTGAAAGATTGTGTGTTTGGTTTTCTTCGACGAGTTATATGTGTAAGAAAGTAATAGAATAACAGTTTTATATAAAAATTCTATGTTGCT   +740

ATATAGATTTATTATCTTAGGATCTGGTATTTGCATGTAGTTGCACAAGCCGG
```

Fig. 2. Nucleotide and amino acid sequence of *hs6871*, *hs6834* and cDNA pFS1968. The arrowhead marks the 5′ end of mRNA as determined by S1 mapping. The brackets mark upstream promoter elements which are at least 70% homologous with the hs consensus sequence CT-GAA-TTC-AG- (Pelham and Bienz, 1982). The potential TATA-box sequence in the promoter region is underlined.

Table I. Codon utilization in soybean *hs6871*

TTT Phe	4	2.6%	TCT Ser	4	2.6%	TAT Tyr	0	0.0%	TGT Cys	0	0.0%
TTC Phe	6	3.9%	TCC Ser	2	1.3%	TAC Tyr	0	0.0%	TGC Cys	0	0.0%
TTA Leu	0	0.0%	TCA Ser	1	0.7%	TAA –	1	–	TGA –	0	–
TTG Leu	2	1.3%	TCG Ser	0	0.0%	TAG –	0	–	TGG Trp	3	2.0%
CTT Leu	2	1.3%	CCT Pro	3	2.0%	CAT His	1	0.7%	CGT Arg	0	0.0%
CTC Leu	2	1.3%	CCC Pro	2	1.3%	CAC His	1	0.7%	CGC Arg	1	0.7%
CTA Leu	0	0.0%	CCA Pro	4	2.6%	CAA Gln	2	1.3%	CGA Arg	3	2.0%
CTG Leu	3	2.0%	CCG Pro	1	0.7%	CAG Gln	1	0.7%	CGG Arg	0	0.0%
ATT Ile	4	2.6%	ACT Thr	2	1.3%	AAT Asn	5	3.3%	AGT Ser	4	2.6%
ATC Ile	1	0.7%	ACC Thr	0	0.0%	AAC Asn	0	0.0%	AGC Ser	4	2.6%
ATA Ile	2	1.3%	ACA Thr	2	1.3%	AAA Lys	2	1.3%	AGA Arg	2	1.3%
ATG Met	2	1.3%	ACG Thr	1	0.7%	AAG Lys	13	8.5%	AGG Arg	4	2.6%
GTT Val	6	3.9%	GCT Ala	4	2.6%	GAT Asp	7	4.6%	GGT Gly	2	1.3%
GTC Val	0	0.0%	GCC Ala	1	0.7%	GAC Asp	5	3.3%	GGC Gly	3	2.0%
GTA Val	1	0.7%	GCA Ala	1	0.7%	GAA Glu	7	4.6%	GGA Gly	1	0.7%
GTG Val	9	5.9%	GCG Ala	1	0.7%	GAG Glu	7	4.6%	GGG Gly	2	1.3%

positions in gene *hs6871* (Figure 2).

The sequence identity of *hs6834* and cDNA 1968 extends beyond the end of the coding region in the 3′ direction to the end of the cDNA. This part of the gene (126 bp) represents an obviously non-translated part of the mRNA. The unexpected lack of a poly(A) stretch at the 3′ end of the cDNA is probably due to S1 nuclease-generated degradation applied in the cDNA cloning protocol or to the observed instability of this cDNA insert in *Escherichia coli* (Schöffl and Key, 1982).

Localization of the 5′ end of mRNA of hs6871

The DNA site corresponding to the 5′ end of the mRNA was determined by DNA/RNA hybridization and S1 nuclease digestion of the hybrids according to Berk and Sharp (1977), as modified by Wasylyk *et al.* (1980). Using the 5′-^{32}P-labelled 208-bp *Taq*I fragment (spanning the DNA stretch between position −162 and +47), hybridization with hs-specific poly(A) RNA was carried out at 42°C prior to S1 digestion. One DNA fragment of 150 ± 2 nucleotides length is protected from S1 digestion by RNA (Figure 3, lane 1). The transcriptional start site, suggested by this shortened fragment, is assigned to the DNA sequence in Figure 2 by an arrow. A weak band, appearing at position 165 in the gel (Figure 3, lane 1) is probably an artefact due to incomplete DNA/RNA denaturation. Other minor bands at position 58–61 in the gel may be explained by cross-hybridization of class I mRNAs (Schöffl and Key, 1982, 1983). High sequence homology within the coding regions of this gene family correlates with the RNA protected 5′ end of *hs6871* up to position −14 in Figure 2, proximal to the translation start.

Discussion

We have determined the nucleotide sequences of one complete (*hs6871*) and one incomplete (*hs6834*) soybean hs gene, both belonging to a multigenic family of low mol. wt. hsps with mol. wts. of 15–18 kd (Schöffl and Key, 1983). The high degree of sequence homology between these two genes confirms their previous classification, which was based on DNA cross-hybridization and hybrid selection and translation of mRNAs. The two genes map ~4 kb apart from each other in direct orientation on a 10.5-kb genomic DNA fragment and are probably both active under hs conditions, as indicated for *hs6834* by identity with the hs-specific cDNA pFS1968 and for *hs6871* by R-loop formation (Schöffl and Key, 1983), S1 mapping and DNA sequencing (this paper). *hs6834* and *hs6871* are closely homologous in their 3′-proximal half of the coding regions. We have unsuccessfully tried to re-isolate a complete copy of *hs6834* from a soybean genomic library cloned into λ-Charon 4A. The partial *Eco*RI digestion used for generating the genebank (Nagao *et al.*, 1981) obviously reduces the possibility of isolating intact genes containing an *Eco*RI site. The soybean hs-gene *hs6871* is the first plant hs gene identified and sequenced at the chromosomal DNA level.

5′-Flanking sequences and the proposed hs promoter

The sequence flanking the 5′ end of the coding region of hs6871 was determined for several hundred nucleotides upstream from the translational start codon. The transcriptional start site was localized ~100 nucleotides upstream from the ATG by S1 mapping. There is one potential 'TATA' or 'Goldberg-Hogness' box (Goldberg, 1978) located 25 nucleotides upstream from the putative mRNA start. A spacing of 20–30 nucleotides between prototypic TATA sequences and the initiation of transcription is a common feature of many eucaryotic genes, which is related to accuracy of transcription (Darnell, 1982; Breathnach and Chambon, 1981; Benoist *et al.*, 1980). This applies probably also for the soybean hs gene *hs6871* with a TAAATA sequence at position −129 to −124 and the transcriptional start at position −103.

The temperature regulation of the transcription of hs genes is probably related to other upstream promoter elements, present in six copies within a DNA stretch of ~130 bp between positions −276 and −149 (see Figure 2). All of these 15 bp long sequences marked in Figure 2 share sequence homologies of 70, 80 or 90% with the 10 bp long hs consensus sequence CT-GAA--TTC-AG- of *Drosophila* (Pelham and Bienz, 1982). As Pelham and Bienz (1982) have shown 80% conservation of the symmetric consensus sequence still allows temperature inducible transcription of linked genes. Although one single copy of the element is sufficient in regulating transcription in *Drosophila* and other animal cells, there are six copies clustered, several of them overlapping each other in the upstream promoter region of the soybean gene. The very effective transcription of this and other low mol. wt. hsp genes (Schöffl and Key, 1982) may be causally related to this accumulation of hs promoter elements. These elements are presumably interacting with as yet unknown factors mediating the hs response (Pelham and Bienz, 1982). In

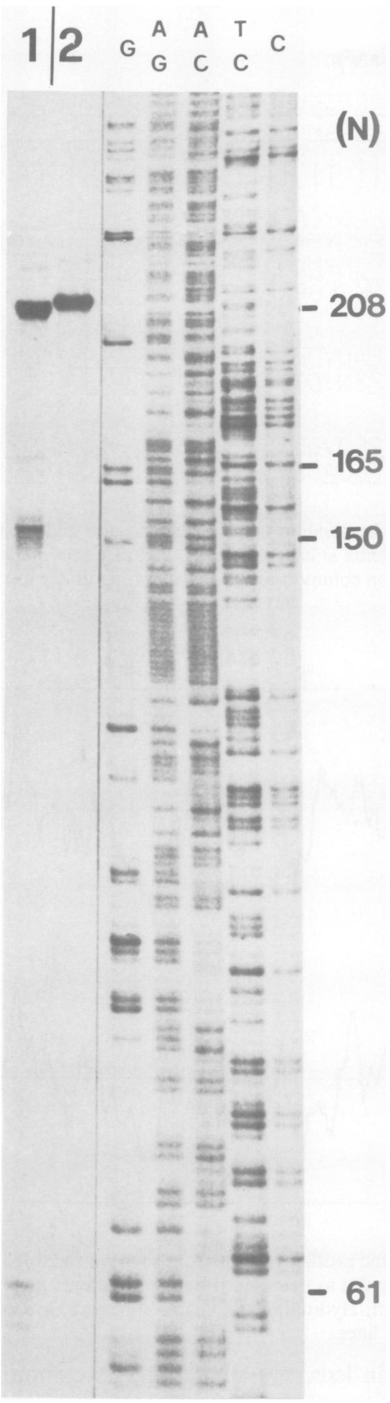

Fig. 3. S1 mapping of the transcription start of *hs6871*. The 208-bp *Taq*I fragment (see Figure 1) was 5′ end-labelled with ^{32}P, hybridized with hs-specific total poly(A) RNA, treated with S1 nuclease and run on a sequencing gel. The protected fragments are displayed in **lane 1**, the untreated fragment is shown in **lane 2**. Lanes G, A+G, A+C, T+C, C represent DNA sequencing tracks used for size determination. Size of fragments is in nucleotides (N).

Drosophila HS Consensus Sequence

5' CTgGAAtnTTCtAGa 3'

Position	Soybean hs6871	Homology
- 276	CccGAAacTTCtAGt	90%
- 245	CcaGAAtgTTtctGa	70%
- 234	CTgaAAgtTTCagaa	70%
- 225	tcaGAAaaTTCtAGt	80%
- 173	CaaGgActTTCtcGa	70%
- 163	CTcGAAagTaCtAta	80%

Proposed Soybean hs6871 Promoter

```
5'..CAAGGACTTTCTCGAAAGTACTATA...17bp...TTTAAATA...20bp...
                                                      |------>
     doublet of overlapping                           mRNA
 'hs consensus promoter elements'        TATA-box     start
```

Fig. 4. DNA sequences of hs promoter elements upstream from soybean *hs6871*. Homology of potential soybean elements refers to the symmetrical consensus sequence (underlined) of *Drosophila* hs genes (Pelham and Bienz, 1982). The elements starting at position −173 and −163 constitute the doublet upstream from the proposed soybean hs promoter.

fact, one of two protein-binding sites in the 5′-flanking region of the *Drosophila* hsp 70 gene covers the upstream hs consensus sequence, the other one the TATA box sequence (Wu, 1984). The most important feature of a soybean hs promoter is probably the connection of a doublet of overlapping hs consensus elements with the TATA-box sequence and the transcriptional start site (see Figure 4). This stretch of DNA, spanning ~70 nucleotides between positions −180 and −110 upstream from *hs6871*, is highly conserved in the potential promoter regions of soybean hs genes (Nagao *et al.*, in preparation). More experimental work is required to prove the functional role of the proposed hs promoter and upstream elements in transcriptional control of plant hs genes.

3′-Flanking regions

The 3′ non-coding extragenic sequences of the soybean hs genes *hs6871* and *hs6834* show a high degree of homology (90%) for ~110−120 nucleotides. However, two minor insertions around position +490 in *hs6834* cause a 7-bp shift of the homologous sequence. The end of homology is only 20 nucleotides shorter than the end of homology with the cDNA, which probably marks the 3′ end of *hs6834*. The 3′ end of *hs6871* is presumably close to that point, not exceeding position +650, since the average size of poly(A) mRNA for class I hsps is 850 nucleotides (Schöffl and Key, 1982) and coding region (460 bases), 5′ non-translated RNA (100 bases) and poly(A) tail (140 bases, according to Key and Silflow, 1975) already add up to 700 nucleotides. We were not able to identify an AATAAA polyadenylation signal (for review, see Proudfoot, 1982) at the end of soybean hs genes. The AATAAA sequence located at position +734 (downstream from *hs6871*) is not conceivably terminating the gene for reasons discussed above. Some other plant genes (soybean leghaemoglobin, maize zein and alcohol dehydrogenase) also do not have AATAAA in the standard position (Jensen *et al.*, 1981; Pederson *et al.*, 1982; Dennis *et al.*, 1984). It is possible that other sequences can act as signals for poly(A) addition in plants.

The 3′-flanking sequence of *hs6871* contains an open reading frame of 135 nucleotides starting in-phase with the just terminated coding region with ATG at position +466 and ending with TAA at position +601. We cannot exclude entirely that a polypeptide is translated from this region, since our protein gels would not have detected such a small protein after *in vitro* translation of the respective mRNAs (Schöffl and Key, 1982; 1983). The low GC content, 38% in contrast to 47% in the coding region, an atypical amino acid composition and probably an inefficient re-initiation of translation

```
Gene/Protein aa-Pos.                   aa-Seque' ce                                   aa-Pos.

Soybean
    hs 6871    80    VLQISGERNVEK--EDKNDTWHRVERSSGKLVRRFRLPENAKVDQVKASMEN-GVLTVTVPK   138
    hs 6834    80    ILQISGERNVEK--EDKNDTWHRVERSSGKFMRSFRLPDNAKVDQVKASMEN-GVLTVTVPK   138

                     ||  | |  |||  |   |  |    ||  |||| |||| || |||| |  | ||||| |||
Drosophila
    hsp 22     80    VLDES-VVLVEAKSEQQEA-EQGGY-SSRHFLGRYVLPDGYEADKVSSSLSDDGVLTISVPN   139
    hsp 23     89    VQDNS--VLVEGNHEERED-DHGFI--TRHFVRRYALPPGYEADKVASTLSSDGVLTIKVPK   146
    hsp 26    108    VVDDS--ILVEGKHEERQD-DHGHI--MRHFVRRYKVPDGYKAEQVVSQLSSDGVLTVSIPK   164
    hsp 27    109    VVDNT--VVVEGKHEERED-GHGHI--QRHFVRKYTLPKGFDPNEVVSTVSSDGVLTLKAPP   165

α-Crystallin
          αB2  93    VLGDV--IEVHGKHEERED-EHGFI--SREFHRRYRLPADVDPLAITSSLSSDGVLTVNGPR   149
          αA2  89    VQEDF--IETHGKHNERQD-DHGYI--SREFHRRYRLPSNVDQSALSCSLSADGMLTFSGPK   145

Caenorhabditis
    hs 1648  (63)    ELDGR-ELKIEGIQEKKS--EHGY--SKRSFSKMILLPEDVDLTSVKSAISNEGKLQIEVPK  (119)
    hs 1641  (32)    KLDGR-ELKIEGIQETKS--EHGY--LKRSFSKMILLPEDADLPSVKSAISNEGKLQIEVPK  ( 98)
```

Fig. 5. Comparison of deduced amino acid sequences of heat-shock genes and bovine α-crystallins. The standard one letter amino acid code is used. Sequences of *Drosophila* hsps (according to Southgate *et al.*, 1983), bovine α-crystallin (Van der Ouderaa *et al.*, 1973, 1974) and of *Caenorhabditis* hsps (deduced from cDNA sequences reported by Russnak *et al.*, 1983) were aligned with the protein region common to the soybean hsps (this paper). Underlined sequences indicate conservation, dashes indicate gaps introduced to obtain best fit of alignment.

due to the unfavorable distal position within the mRNA (according to Kozak, 1984), makes it unlikely that a second polypeptide is translated from the same RNA.

Protein-coding sequences

The amino acid sequences of the proteins encoded by *hs6871* and *hs6834* share 92% homology. 17 out of 20 nucleotide changes in the DNA are in the 3rd position of codons and 16 of them do not change the amino acid sequence. The amino acid composition of these soybean hsps shows a typical high lysine content (see Table I) and a preference for the lysine codon AAG (8.5%) over AAA (1.3%). *Drosophila* hs genes have a similar bias for the AAA codon as deduced from the data of hsp 70 (Ingolia *et al.*, 1980) and of hsp 22, 23, 26 and 27 (Southgate *et al.*, 1983). In contrast, a balanced usage of lysine codons can be deduced from the DNA sequences of other soybean genes as for example conglycinine (Schuler *et al.*, 1982), leghaemoglobin (Brisson and Verma, 1982) and lectin genes (Vodkin *et al.*, 1983). This difference in codon utilization for lysine between hsps and non-hsps correlates with a translational preference of hs-specific mRNAs under hs conditions (Key *et al.*, 1981). A hs-dependent activation or preservation of tRNAs for prevalent amino acids could be essential for the synthesis of hsps. In this context it is remarkable that one of the *E. coli* hsps is a tRNA^{lys} synthetase (Neidhardt *et al.*, 1982).

The soybean hsps described in this paper share a homologous stretch of amino acids in the central and carboxy-terminal region with hsps from other organisms and with bovine α-crystallin (see Figure 5). The order of sequences compared with the soybean hsps in Figure 5 reflects decreasing homology. However, the hydropathic profiles (Kyte and Doolittle, 1982) of these proteins all show the conservation of structural domains as demonstrated in Figure 6 for soybean *hs6871* and *Drosophila* hsp 22. The most striking features are a hydrophobic peak, including the very conserved GVLT pattern of amino acids located at position 130 – 134 in *hs6871* (Figures 5 and 6) and the very large hydrophilic domain spread over ~30 amino acids from the center of the protein towards its carboxy-terminal end. This hydrophilic region is probably α-helical (Southgate *et al.*, 1983) according to secondary structure predictions (Garnier *et al.*, 1978).

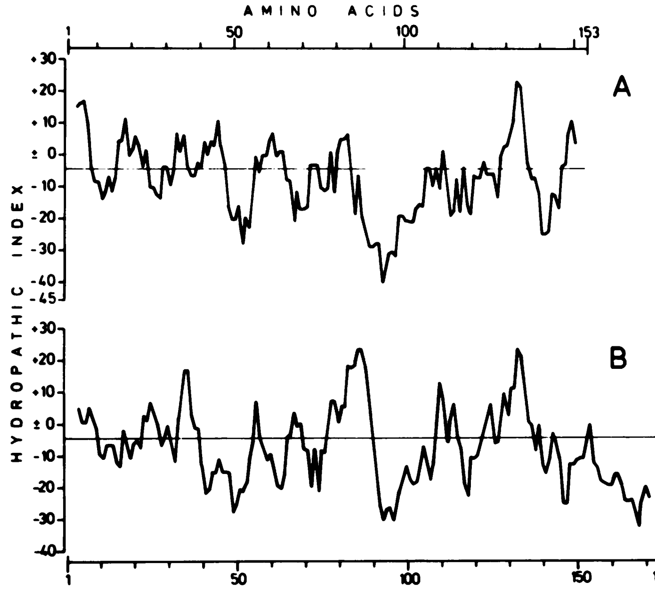

Fig. 6. Hydropathic profiles of hsps from soybean (**A**:*hs6871*) and *Drosophila* (**B**:hsp 22) as determined by the Kyte and Doolittle (1982) computer program. Hydrophobic domains are above, hydrophilic domains below the middle lines.

The resemblance between structural features of hsps and the mammalian lens crystallins suggests common functional properties. This could be protein aggregation, which may occur in soybean in a temperature-dependent fashion, when low mol. wt. hsps associate with nuclei and cell organelles exclusively at high temperature (Key *et al.*, 1982; Lin *et al.*, 1984; Schöffl *et al.*, 1984). The nature of an association between hsps and other cellular components is unknown, but their short hydrophobic amino-terminal ends (see Figure 6) may perhaps interact with membranes during the hs response.

Materials and methods

DNA sequencing

DNA sequencing was done chemically according to Maxam and Gilbert (1980) with the modifications of Volkaert *et al.* (1984). DNA fragments of interest were subcloned into the newly constructed sequencing vector pSVB20 (Arnold, personal communication). This vector is a derivative of plasmid pUC8 (Vieira and Messing, 1982) and contains in addition a single *Bst*E II site (5′ G/GTCACC) within the multiple cloning site. The additional restriction site does not disrupt the reading frame of the surrounding *lacZ* protein and allows rapid single end-labelling of the DNA by a 3′ filling in reaction with DNA polymerase I/Klenow fragment, dGTP, dTTP and [α-^{32}P]dCTP. Subclones (see Figure 1) were maintained in the *E. coli* strain JM83 (Vieira and Messing, 1982). Plasmid DNA was isolated by the sarcosyl lysis method (Bazaral and Helinski, 1968), purified twice by CsCl ethidium bromide density gradient centrifugation, cut with *Bst*E II, labelled at a single end and sequenced. Sequencing reactions were separated on 0.3 mm thin 5% polyacrylamide gels with the LKB Macrophor gelsystem. The Pustell and Kafatos (1982) computer program was used for portions of the nucleotide sequence analysis; the Kyte and Doolittle (1982) program was used for the determination of hydropathic profiles of proteins.

5′ S1 mapping

The 5′ ends of *hs6871* mRNA were determined by S1 mapping (Berk and Sharp, 1977), as modified by Wasylyk *et al.* (1980). Dephosphorylated *Taq*I fragments were 5′ end-labelled with [γ-^{32}P]ATP, isolated from polyacrylamide gels according to Maniatis *et al.* (1982), hybridized for 4 h to total hs poly(A) RNA from soybean hypocotyl according to Key *et al.* (1981) and treated with nuclease S1 (1000 U/ml) for 30 min at 37°C. Protected fragments were sized on a 5% sequencing gel using G, A + G, A + C, T + C, C, sequencing tracks of a known DNA sequence as a reference.

Acknowledgements

We thank Dr. J.L.Key, University of Georgia, Athens, USA for helpful discussions and his support of our collaboration, Dr. M.O'Connel and M.Hynes for critical reading, Hanne Schöffl for typing of the manuscript and Sieglinde Angermüller for technical assistance. We are indebted to W.Arnold, Universität Bielefeld, who kindly provided the as yet unpublished sequencing vector pSVB20. We also greatly appreciate a computer search for protein homologies and hydropathic profiles of hsp 22 and *hs6871* by Dr. R.F.Doolittle, University of California, San Diego, USA. This work was supported by grant Scho 242-4 of the Deutsche Forschungsgemeinschaft. R.T.N. was supported by a contract from Agrigenetics Res. Assoc., limited to J.L.Key.

References

Ashburner,M. and Bonner,J.J. (1979) *Cell*, **17**, 241-254.

Bazaral,M. and Helinski,D.R. (1968) *J. Mol. Biol.*, **36**, 185-194.

Benoist,C., O'Hare,K.O., Breathnach,R. and Chambon,P. (1980) *Nucleic Acids Res.*, **8**, 127-142.

Berk,J.B. and Sharp,P.A. (1977) *Cell*, **12**, 721-732.

Bienz,M. and Pelham,H.R.B. (1982) *EMBO J.*, **1**, 1583-1588.

Breathnach,R. and Chambon,P. (1981) *Annu. Rev. Biochem.*, **51**, 813-848.

Brisson,N. and Verma,D.P. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 4055-4059.

Burke,J.F. and Ish-Horowicz,D. (1982) *Nucleic Acids Res.*, **10**, 3821-3830.

Corces,V., Holmgren,R., Freund,R., Morimoto,R. and Meselson,M. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 5390-5393.

Corces,V., Pellicer,A., Axel,R. and Meselson,M. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 7038-7042.

Darnell,J.E. (1982) *Nature*, **297**, 365-371.

Dennis,E.S., Gerlach,W.L., Pryor,A.J., Bennetzen,J.L., Inglis,A., Llewellyn,D., Sachs,M., Ferl,R.J. and Peacock,W.J. (1984) *Nucleic Acids Res.*, **12**, 3983-3999.

Garnier,J., Osguthorpe,D.J. and Robson,B. (1978) *J. Mol. Biol.*, **120**, 97-120.

Goldberg,M. (1978) Dissertation, Stanford University, USA.

Ingolia,T.D. and Craig,E.E. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 2360-2364.

Ingolia,T.D., Craig,E.A. and McCarthy,B.J. (1980) *Cell*, **21**, 669-679.

Jensen,E.O., Palendan,K., Hyldig-Nielson,J.J., Jorgenson,P. and Marcker, K.A. (1981) *Nature*, **291**, 677-679.

Karch,F., Torok,J. and Tissières,A. (1981) *J. Mol. Biol.*, **148**, 219-230.

Key,J.L. and Silflow,C. (1975) *Plant Physiol.*, **56**, 364-369.

Key,J.L., Lin,C.Y. and Chen,Y.M. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 3526-3530.

Key,J.L., Lin,C.Y., Ceglarz,E. and Schöffl,F. (1982) in Schlessinger,M.J., Ashburner,M. and Tissières,A. (eds.) *Heat Shock, From Bacteria to Man*, Cold Spring Harbor Laboratory Press, NY, pp. 329-336.

Key,J.L., Czarnecka,E., Lin,C.Y., Kimpel,J., Mothershed,C. and Schöffl, F. (1983) in Randall,D.D., Blevins,D.G., Larson,R.L. and Rapp,B.J. (eds.), *Current Topics in Plant Biochemistry and Physiology*, University of Missouri, Columbia, MO, pp. 107-118.

Key,J.L., Kimpel,J., Vierling,E., Lin,C.Y., Nagao,R.T., Czarnecka,E. and Schöffl,F. (1984) in Atkinson,B. and Walden,D.B. (eds.), *Changes in Gene Expression in Response to Heat Shock*, Academic Press, in press.

Kozak,M. (1984) *Nucleic Acids Res.*, **12**, 3873-3893.

Kyte,J. and Doolittle,R.F. (1982) *J. Mol. Biol.*, **157**, 105-132.

Lin,C.Y., Roberts,J. and Key,J.L. (1984) *Plant. Physiol.*, **74**, 152-160.

Maniatis,T., Fritsch,E.F. and Sambrook,J., eds. (1982) *Molecular Cloning. A Laboratory Manual*, published by Cold Spring Harbor Laboratory Press, NY.

Maxam,A. and Gilbert,W. (1980) *Methods Enzymol.*, **65**, 499-560.

Mirault,M.E., Delwart,E. and Southgate,R. (1982) in Schlesinger,M., Ashburner,M. and Tissières,A. (eds.), *Heat Shock From Bacteria to Man*, Cold Spring Harbor Laboratory Press, NY, pp. 244-248.

Nagao,R.T., Shah,D.M., Eckenrode,V.K. and Meagher,R.B. (1981) *DNA*, **2**, 1-9.

Neidhardt,F.C., Van Bogelen,R.A. and Lau,E.T. (1982) in Schlesinger,M., Ashburner,M. and Tissières,A. (eds.), *Heat Shock From Bacteria to Man*, Cold Spring Harbor Laboratory Press, NY, pp. 131-138.

Pederson,K., Devereux,J., Wilson,D.R., Sheldon,E. and Larkins,B.A. (1982) *Cell*, **29**, 1015-1026.

Pelham,H.R.B. (1982) *Cell*, **30**, 517-528.

Pelham,H.R.B. and Bienz,M. (1982) *EMBO J.*, **1**, 1473-1477.

Proudfoot,N. (1982) *Nature*, **298**, 516-517.

Pustell,J.M. and Kafatos,F.C. (1982) *Nucleic Acids Res.*, **10**, 51-59.

Russnak,R.H., Jones,D., Peter,E. and Candido,M. (1983) *Nucleic Acids Res.*, **11**, 3187-3205.

Schlesinger,M.J., Ashburner,M. and Tissières,A. eds. (1982) *Heat Shock From Bacteria to Man*, published by Cold Spring Harbor Laboratory Press, NY.

Schöffl,F. (1984) in *Gentechnik, Dechma Monograpie*, Vol. **95**, Verlag Chemie, Weinheim, pp. 323-337.

Schöffl,F. and Key,J.L. (1982) *J. Mol. Appl. Genet.*, **1**, 301-314.

Schöffl,F. and Key,J.L. (1983) *Plant. Mol. Biol.*, **2**, 269-278.

Schöffl,F., Lin,C.Y. and Key,J.L. (1984) in Stewart,G.R. and Lea,P.J. (eds.), *Genetic Manipulation of Plants and its Application to Agriculture*, Vol. **23**, Oxford University Press, in press.

Schuler,M.A., Schmitt,E.S. and Beachy,R.N. (1982) *Nucleic Acids Res.*, **10**, 8225-8240.

Southern,E.M. (1975) *J. Mol. Biol.*, **98**, 503-517.

Southgate,R., Ayme,A. and Voellmy,R. (1983) *J. Mol. Biol.*, **165**, 35-57.

Van der Ouderaa,F.J., de Jong,W.W. and Bloemendal,H. (1973) *Eur. J. Biochem.*, **39**, 207-211.

Van der Ouderaa,F.J., de Jong,W.W. and Bloemendal,H. (1974) *Eur. J. Biochem.*, **49**, 157-161.

Velasquez,J.M. and Lindquist,S. (1984) *Cell*, **36**, 655-662.

Vieira,J. and Messing,J. (1982) *Gene*, **19**, 259-268.

Vodkin,L.O., Rhodes,P.R. and Goldberg,R.B. (1983) *Cell*, **34**, 1023-1031.

Voellmy,R., Goldschmidt-Clermont,M., Southgate,R., Tissières,A., Levis,R. and Gering,W. (1981) *Cell*, **23**, 261-270.

Voellmy,R. and Runngger,D. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 1776-1780.

Volkaert,G., Winter,G. and Gaillard,C. (1984) in Pühler,A. and Timmis, K.N. (eds.), *Advanced Molecular Genetics*, Springer Verlag, Berlin, pp. 249-280.

Wadsworth,S.C., Craig,E.A. and McCarthy,B.J. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 2134-2137.

Wasylyk,B., Kedinger,C., Corden,J., Brison,O. and Chambon,P. (1980) *Nature*, **285**, 367-373.

Wu,C. (1984) *Nature*, **309**, 229-234.

Received on 6 July 1984

Note added in proof

The transcription of *hs6871* has been confirmed in our laboratory by studies on the expression of this gene after transfer into a heterologous genetic background (Schöffl and Baumann, manuscript in preparation).